Heart Disease Prediction Using Machine Learning

- by Arshadul Shaikh

Summary:

In this project, we have successfully built a **Machine Learning model** that will keep getting better as we provide more data to predict the likelihood of heart disease within **10 years**.
After testing multiple models, Logistic Regression (Baseline Model), Decision Tree Classifier (Interpretable), Random Forest Classifier with and without tuned HyperParameters, XGBClassifier, Support Vector Machine (SVM), & Neural Network (Feedforward) (Deep Learning Approach) we finalized **Random Forest Classifier** as our best-performing model with the accuracy of 77%.

Even though Random Forest had an improved accuracy, all models struggled to have the good Precision, recall and F1-score. This meant all models were biased towards predicting "No Heart Disease" (class 0) this would have been true, if we had not done resampling to have equal population of class 1 & class 0 but with sampling as well we struggled to have good metrics even after tuning hyperparameters. XGBClassifier which uses **Gradient Boosting**, meaning it **focuses on mistakes** from previous trees as well didn't managed to get a good result.

This was seen because we had several features which made the model complex, and also some of the important feature like **family history of heart disease** as a factor, **exercise & dietary habits** could have increased our model overall.

In the end, we have the model function ready to be used by medical team by feeding inputs to predict the possibilities and hence create awareness among patients, so they can rectify and make lifestyle changes for a better living.

Introduction:

Heart disease continues to be the leading cause of mortality both globally and in the United States. According to the American Heart Association's 2025 Heart Disease and Stroke Statistics Update, cardiovascular diseases (CVDs) remain the foremost cause of death worldwide.

In 2021, CVDs accounted for approximately 32% of all global deaths, with ischemic heart disease and stroke being the predominant contributors.

In 2022, coronary heart disease (CHD) was responsible for 39.5% of all CVD-related deaths in the U.S., followed by stroke (17.6%), other CVDs (17.0%), hypertensive diseases (14.0%), heart failure (9.3%), and diseases of the arteries (2.6%)

Risk Factors and Trends:

- The prevalence of key risk factors such as hypertension, obesity, and diabetes continues to rise, contributing significantly to the burden of heart disease. Notably, over half of U.S. adults are affected by diabetes or prediabetes

Gender-Specific Data:

- Heart disease affects more than 60 million women in the United States, underscoring the critical need for targeted prevention and treatment strategies.

I have tried to work on the Framingham dataset which is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients' information. It includes over 4,000 records and 15 attributes. Variables Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

Understanding Dataset:

Attributes of our Dataset: I have divided the columns details as per their domain for having better understanding

- Demographic:
  - Sex: male or female(Nominal)
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
  - Education: no further information provided
- Behavioral:
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- Information on medical history:
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
- Information on current medical condition:
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)

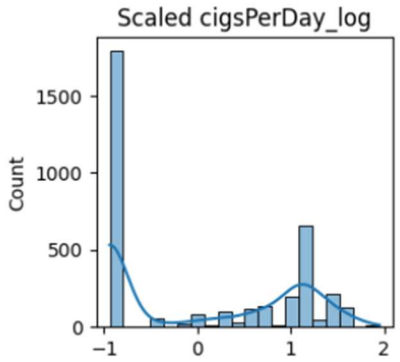
    - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
    - Glucose: glucose level (Continuous)
- Target variable to predict:
    - 10 year risk of coronary heart disease (CHD) - (binary: “1”, means “Yes”, “0” means “No”)

Analysis:

We have successfully built a Machine Learning model to predict the likelihood of heart disease within 10 years.

Key EDA Insights

- **No Major Outliers:** Most numerical variables follow reasonable distributions.
- **Skewed Features:** cigsPerDay was right-skewed, requiring transformation (log transformation used), which was log transformed and then scaled.



- **Imbalanced Target Variable:**
    - **91%** of individuals **did not** develop heart disease (TenYearCHD = 0).
    - **Only 9%** developed heart disease (TenYearCHD = 1).
    - This required **oversampling (SMOTE)** to balance the dataset.

Data Preprocessing Steps

- **Handling Missing Values:**
    - education, cigsPerDay, BPMeds, totChol, BMI, heartRate, and glucose had missing values.
    - We **imputed low-missing features i.e. education**, **dropped high-missing glucose values** to avoid bias. BPMeds (Blood Pressure Medication) is a medical indicator,

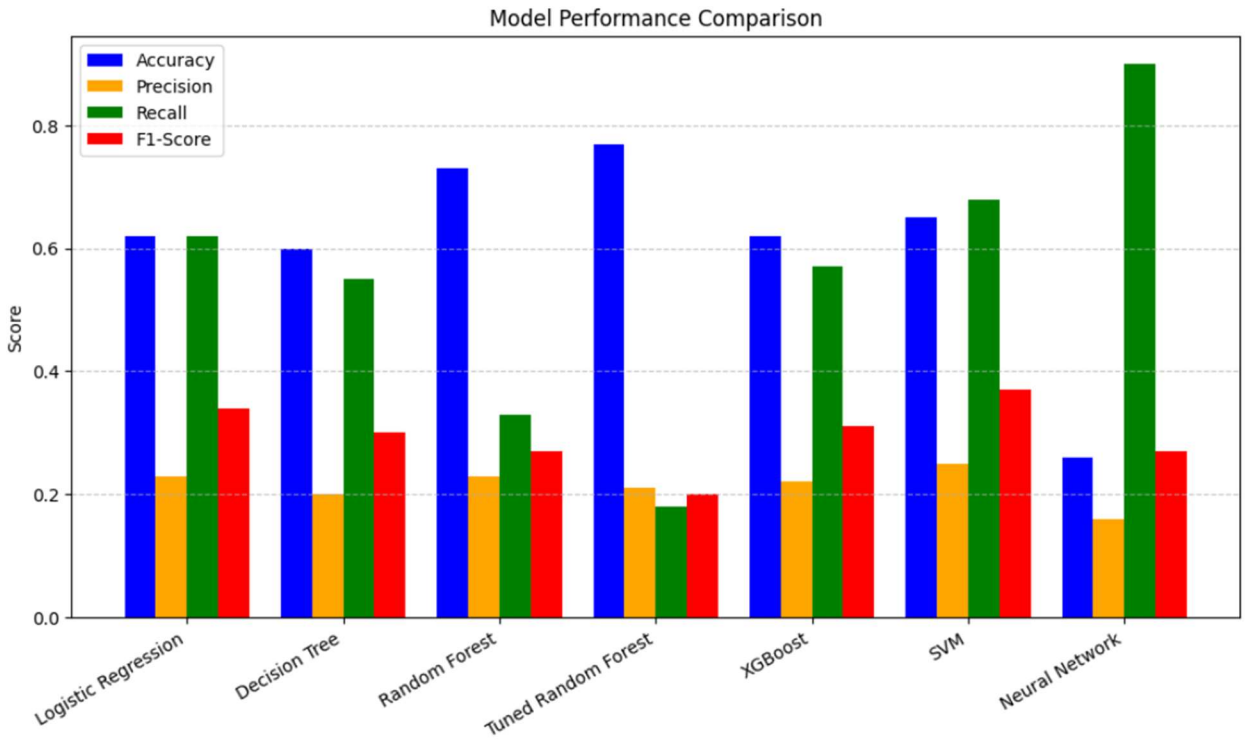
simply filling missing values with the most frequent value (mode) may not be the best approach, hence used regression models to predict the right values.

- **Feature Engineering:**
    - Created cigsPerDay_log to handle skewed smoking data.
    - Stopped considering non-informative features like education.
- **Scaling & Transformation:**
    - Applied **StandardScaler** to normalize numerical variables.
    - Categorical features were already encoded as **binary (0/1)**.
- **Balancing the Dataset:**
    - Used **SMOTE (Synthetic Minority Oversampling)** to oversample the minority class (patients with heart disease).
    - This ensured that the model learned **both classes equally**.

Machine Learning Models Implemented:

We trained several machine learning models and **compared their performance**:

Model	Accuracy	Precision (Class 1)	Recall (Class 1)	F1-Score (Class 1)
Logistic Regression	62%	23%	62%	34%
Decision Tree	60%	20%	55%	30%
Random Forest	73%	23%	33%	27%
Tuned Random Forest	**77%**	**21%**	**18%**	**20%**
XGBoost	62%	22%	57%	31%
SVM	65%	25%	68%	37%
Neural Network (MLP)	26%	16%	90%	27%



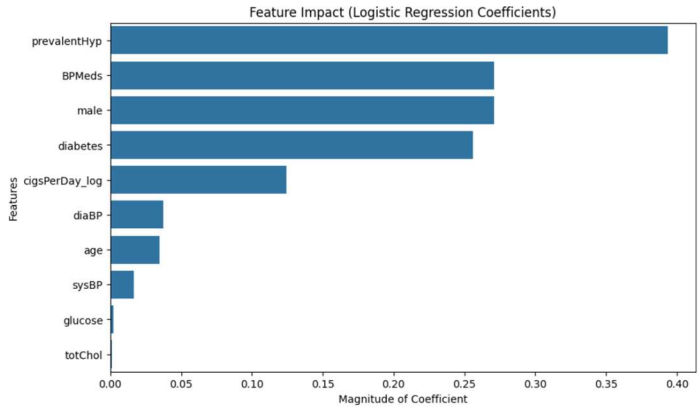
Model Selection & Performance Evaluation

- **Random Forest (Tuned) was the best model** with **77% accuracy**, outperforming others.
- **However, recall (21%) for class 1 (heart disease cases) was still low**, meaning many actual cases were missed.
- **Other models like SVM & XGBoost** had better recall but lower overall accuracy.
- **Neural Network performed poorly**, likely due to limited data.
- Overall **Random Forest (not tuned)** was the selected based on Precision, Recall, F1 score.

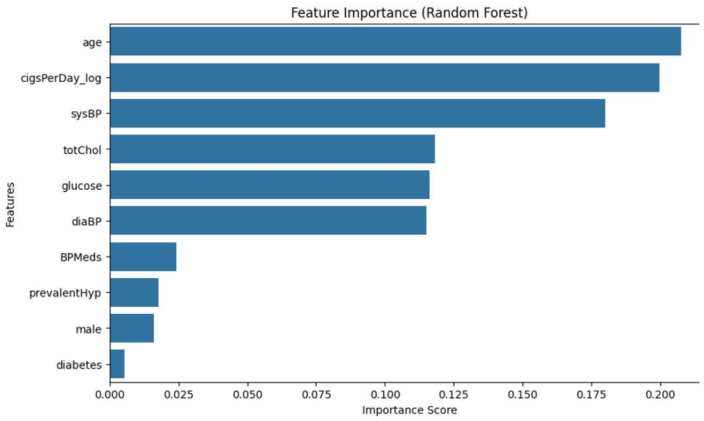
Feature Importance Analysis

We analyzed which features contributed most to predictions:

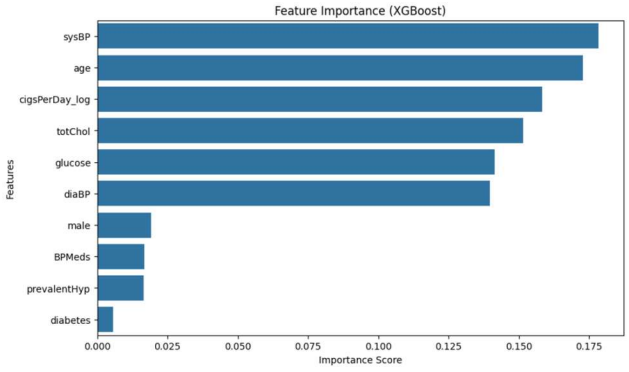
- **Logistic Regression:**
    - **Age, Diabetes, Systolic BP, Prevalent Hypertension** were most important.



- **Random Forest:**
    - **Systolic BP, Age, Cholesterol, BMI** had the highest impact.



- **XGBoost:**
    - **Glucose, Age, Systolic BP, Diabetes** were key predictors.



- **Neural Networks (Permutation Importance):**
    - **Age, BMI, Glucose** were top contributors.

Final Thoughts & Future Improvements

Strengths of the Model:

- **Random Forest** was the **best-performing model** with the highest accuracy.
- **Feature Importance Insights** confirmed medical relevance (e.g., BP, Diabetes, Cholesterol).
- **SMOTE** helped improve model learning on imbalanced data.

Limitations & Next Steps:

- **Low Recall for Heart Disease Cases (Class 1)** → we would need more data to make the model robust.
- **Cost-sensitive Learning** → Penalize **false negatives more heavily**.
- **More Data & Feature Engineering** → Investigate **new predictors** (e.g., lifestyle, genetic factors).
- **Deploy as a Web API** → Allow doctors to input patient data & receive risk predictions.

This project demonstrated the **power of machine learning in medical diagnostics**, but further improvements are needed before clinical adoption.

CITATIONS & References referred:

- Kaggle repository used:
  - https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset
  - https://pubmed.ncbi.nlm.nih.gov/14819398/

- Machine Learning Models & Techniques: Logistic Regression, Decision Trees, Random Forest, XGBoost, SVM, Neural Networks (MLP)
  - Source: Scikit-Learn & XGBoost Documentation

References:

Scikit-learn: Machine Learning in Python.
- https://scikit-learn.org/stable/supervised_learning.html
- https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
- https://scikit-learn.org/stable/modules/ensemble.html
- https://arxiv.org/abs/1603.02754
- https://scikit-learn.org/stable/modules/svm.html

Performance Evaluation Metrics.

- Accuracy, Precision, Recall, F1-Score, ROC Curve

Source: Scikit-Learn Documentation
https://scikit-learn.org/stable/model_selection.html

Reference:
https://pubmed.ncbi.nlm.nih.gov/25738806/

- SMOTE (Synthetic Minority Over-sampling Technique) for Imbalanced Data
  - Source: imbalanced-learn Documentation

    - https://imbalanced-learn.org/stable/over_sampling.html

    Reference:
    - https://www.jair.org/index.php/jair/article/view/10302

- Cost-Sensitive Learning & Threshold Calibration
  - Adjusting Classification Thresholds, Class Weighting in Random Forest

    Source: Scikit-Learn Documentation
    Reference:
    - https://cseweb.ucsd.edu/~elkan/rescale.pdf

- Python Libraries Used:
  - Scikit-Learn, Pandas, NumPy, Matplotlib, Seaborn, imblearn.